DOCUMENT RESUME

ED 096 840 FL 006 487

AUTHOR	Wrenn, James J.
TITLE	A Standard Sample of Present-Day Chinese for Use with Digital Computers. Final Report.
INSTITUTION	Brown Univ., Providence, R.I.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
BUREAU NO	BR-0-7756
PUB DATE	May 74
CONTRACT	OEC-0-70-4544(823)
NOTE	13p.
AVAILABLE FROM	Manual and Corpus available from the Department of Linguistics, Brown University, Providence, Rhode Island 02912 ($50.00 if tapes are furnished, $75.00 otherwise)
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	Chinese; Codification; *Computational Linguistics; *Data Bases; Data Collection; Digital Computers; Information Retrieval; *Mandarin Chinese; Mathematical Linguistics; Vocabulary; Word Frequency; *Word Lists
IDENTIFIERS	NDEA Title VI

ABSTRACT

The final report on a project to develop a standard corpus of present-day Mandarin Chinese is presented. This corpus consists of words of running text of Chinese prose printed in the Republic of China during the calendar year 1968. The corpus, although originally planned to have a total of 500 samples of 2000 words each, has only 294 samples. Each sample starts at the beginning of a sentence, but not necessarily at the beginning of a paragraph or larger division. The samples represent a variety of styles of modern prose, selected for their representative quality rather than their literary merit. The collection consists primarily of samples from books and some major periodicals available through the library at the National Taiwan University and the National Central Library. For each sample collected, a copy was made and then transcribed into a modified Pin-yin romanization. For each sample, counts were taken of the following: names, formulae, figures, foreign strings, foreign words, words (in total), and syllables. After the samples were collected and romanized, they were then codified. A manual accompanies the corpus, which comprises one magnetic tape of about 1,200 feet, available in either 7-track or 9-track mode. (Author/LG)

Final Report

Contract No. OEC-0-70-4544 (823)

James J. Wrenn
Brown University
Providence, Rhode Island 02912

A STANDARD SAMPLE OF PRESENT-DAY CHINESE FOR USE WITH DIGITAL COMPUTERS

May 1974

The research reported herein was supported by the
Office of Education,
U.S. Department of Health, Education, and Welfare
under the authority of
Title VI, Section 602 of the
National Defense Education Act

Final Report

Contract No. OEC-0-70-4544 (823)

A STANDARD SAMPLE OF PRESENT-DAY CHINESE FOR USE WITH DIGITAL COMPUTERS

James J. Wrenn

Brown University
Providence, Rhode Island 02912

May 1974

The research reported herein was performed persuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education

Contents

This standard corpus of present-day Chinese consists of words of running text of Chinese prose printed in the Republic of China during the calendar year 1968. Some of this material has undoubtedly been written earlier, but no material known to be a second edition or reprint of an earlier text has been used.

The corpus, although originally planned to have a total of five hundred samples of 2000 words each, and fully comparable to the corpus of American English collected by Professor W. Nelson Francis in 1964, has only 294 samples. Each sample begins at the beginning of a sentence, but not necessarily at the beginning of a paragraph or larger division. The samples represent a variety of styles and varieties of modern prose. As with the English corpus, verse was not included on the ground that it presents special linguistic problems different from those of prose. Drama was excluded for similar reasons. Fiction was included, but samples with more than 50% of spoken dialogue were excluded, as were those which consisted of more than 50% Literary Chinese (wen2-yan2). Samples were chosen for their representative quality rather than for their literary merit or other qualities.

The project was begun with a conference on the collection of a corpus held at Brown University on 30 June 1970. Here an attempt was made to establish the categories and sub-categories of the Chinese Corpus as contrasted with the English corpus, and a variety of special details on the collection of the corpus were specified. In addition, there were a number of specific suggestions on the coding of input, and on difficulties which should be avoided.

On arrival in Taiwan, and based on suggestions from the conference, I made an effort to adjust any implicit cultural bias of the preliminary list of categories and the numbers of samples to be included from each, and to adjust more accurately to the publishing situation in Taiwan by consulting the publications directory for the province of Taiwan for the previous year (Chu1-ban3 Shi4-ye4 Deng1-ji4 Yi1 Lan2, Taipei, Ministry of the Interior, 1968), and the acquisitions lists of the National Central Library for a period of approximately two and one-half years. I prepared acquisitions statistics for all the books and periodicals, separated according to

BEST COPY AVAILABLE

the Dewey Decimal Classification System. The statistics for acquisitions in all of these categories were averaged, and computed to a percentage of a number of samples in a 500 word sample corpus.

Then, these numbers were used as guidelines for the final collection, constituting limits on the number of samples in each category or sub-category.

The universe from which the samples were chosen was further defined as selections from works held in the Library of the National Taiwan University, and when the holdings of this library proved inadequate or inaccessible, further supplemented from the holdings of the National Central Library, and occasionally from other sources in the case of ephemera no longer available in the larger libraries.

Generally, we began with the collection of the samples from books, and some major periodicals which were available to us through the Library of the National Taiwan University, attempting to collect samples in each category. The first step in assuring randomness at this stage was the preparation of a slip for each book published in 1967 held by the library, with call number, author and title. Random number tables were then used in selecting the book by its call number, treating the call number as a series of digits, and selecting smaller and smaller subsets of the whole file with each look-up until only a single title was left. The sample was defined using similar procedures with the page numbers.

Unfortunately there proved to be no easy way to pre-count the words so as to assure that there would be 2000 in each sample, and various expedients were tried, all imprecise. Therefore, many of the samples fall short of the 2000-word goal, with those collected earlier farthest from the mark, where we had used the number of characters in the sample as guide. Later, with experience, we learned to have the transcribers count the total number of lines of text transcribed instead, with consistently better results.

For each sample selected, a copy was made on locally available copying equipment, since many of the materials were not circulable, and the copy was given to a typist for transcription into a modified Pin-yin romanization. Romanization was chosen as the vehicle for transcription only partly because of its increasing acceptance within the Chinese culture.

The most important reason for the choice of this

BEST COPY AVAILABLE

romanization rested in the assumption that the text would be used by Chinese and Westerners familiar with Chinese, and that they would be more comfortable reading romanized text than reading material transcribed into one of the versions of the telegraphic code, or into some specialized machine coding. This choice was made with the awareness that some ambiguity would remain because different words would be spelled in the same way. We assumed that the literate user could disambiguate such forms from the context, and made special efforts in the coding to minimize homonymy.

The samples were first transcribed, the transciption was proof-read, then the proof-read and edited copy was keypunched, and the cards checked for errors.

After the cards had been read onto tape, the sample was again checked by machine with a simple program designed to check for the two commonest errors, absence of a tone-mark on any syllable, and the insertion of a space in the middle of a word, caused when the keypunch operator left a space at the end of the card at the end of a syllable that did not finish a word. When these were found, corrections were made by inserting a new, corrected, card.

For each sample, counts were made for the following items: Names, undifferentiated as to whether they are personal or place names; Formulae; Figures; Foreign Strings, which are separate occurrences of one or more foreign words; and Foreign Words, the total of the foreign words listed in the separate strings; Words, a total of all of the separate graphically defined words; followed by a listing of total syllables in the sample, with two subsets of this last: one being syllables preceded by a plus sign; and the other syllables preceded by an asterisk, which distinctions have grammatical significance, as noted in the section on Coding Conventions, below. For each sample, the list of counts follows the bibliographic data.

In May, 1974 the total counts are: Names, 25,928; Formulae, 69; Figures, 80; Foreign Strings, 2,137; Foreign Words, 6,419; Words, 965,536; Syllables, 1,070,932; +Syllables, 86,037; *Syllables, 4,853.

The coding procedures that allow retrieval of the many classes of items such as names and formulae introduce some features that are inconsistent with the requirement for the unity of the graphic word. For example, the modification particle '-de5', should be immediately linked

BEST COPY AVAILABLE

to the previous syllable. However, when it is preceded by a word which is a personal or place name, coding conventions come into conflict, and the requirement that the symbols which designate a name be separated from surrounding text by a blank space takes precedence in the ordered set of coding rules. Since some of these may be potentially confusing to the user, two separate lists have been prepared which reflect two different levels of specificity with respect to the scale of inconsistency.

The first, which, according to more economical standards of definition, may reflect coding errors at the more extreme end of the inconsistency scale is appended directly to the bibliographic description of the sample itself, and is headed 'Possible Errors'. This list is presented in two columns in the text of the manual, with the column to the left representing the offending item, and the context in which it occurs shown in the column to the right, with the context displayed with ten spaces to the left and right of the item, while the item itself is represented by an underscore. The ten-space limitation is usually effective in delimiting the context so that the user can judge whether a real error has been made, but also usually offends the eye and the sense in that it breaks the majority of both preceding and following words in such a way as to be aesthetically displeasing. Many of the items in these lists represent faithful recognition of repetitive errors not taken full account of in the initial coding. Specifically, of the items listed as 'Possible Errors' for the sample A03, most were collected because the programs we used specified that a syllable not followed by a numeral indicating a full or neutral tone was to be considered as a 'Possible Error'. The list there presented was machine selected because the letters 'A', 'B', 'C', and 'D' in the combinations 'Group A', 'Group B', 'Group C', and 'Group D' were, properly, not followed by tone-indicating digits.

Similarly, the long listing of 'Possible Errors' for sample G34 reflects assiduous machine concern with the fact that similar lack of tone tagging is reflected on each repetition of each letter in the frequent citations in this sample of the drug LSD. The human editor can deal with this anomaly lightly and dismiss it as of no concern, but lack of such formalized machine procedures may cause the user to overlook such real errors as that in A09 or the many in J02 in which the omission of a space preceding or following an item may well prevent its inclusion in a

BEST COPY AVAILABLE

concordance until it has been corrected. On the whole it seemed better to leave this level of 'Possible Error' associated directly with the text and its description.

For the second list, which seemed on inspection to represent a lower level of real error, and a higher incidence of coding anomalies, or of coding rules in conflict with respect to their relative rank, it seemed more efficient to relegate the much longer list to an appendix listed according to sample, and it is there so presented.

Problems

Initially it was found that there were very few individuals who were willing to work on a short-term project of such magnitude. Additionally there was a problem of finding individuals whose typing skills were adequate for rapid conversion from Chinese characters into romanized text and whose dialect was sufficiently standard to assure that they could represent Chinese in a standard romanization. Dialect variance was shockingly wide-ranging. As an initial partial solution, a single typist was hired as a full-time secretary-transcriber. She had had considerable experience in English typing, spoke standard Mandarin, and had worked with romanized Mandarin previously. Over an initial few months it developed that there were both personal and institutional problems in requiring individuals to do careful transcription over long periods during any working day. As the problems became clearer, it became apparent that the intensity involved in detailed, accurate transcription of technical and literary material of this type could not be done full-time by any single individual for very many hours per day. We experimented briefly with two or three part-time typists, who were personally recommended for their skills, but found both quality and output to be low. At this point we began the search for competent, experienced, careful, quality-oriented typist-transcribers who could handle the range of material that we had chosen. It turned out that no one was interested in working full-time and that we should put our efforts into recruiting a larger number of part-time typist-transcribers. By this time the first 6 months of the initial contract period had passed with only limited production. It was clear that any attempt to use a large

BEST COPY AVAILABLE

number of part-time typist-transcribers would require much more clearly described definition of tasks, expectations, and standards for work than had so far been described. Therefore, I turned my efforts to developing a programmed outline to teach transcription from Chinese characters into Romanized Chinese in the Pin-yin system. The program begins with simple consonants, and increases in complexity, progressing to vowels, syllables, the tonal notation adapted for the input of of unstressed syllables; then to the techniques for joining syllables into words, and for indicating bound forms and similar grammatical adjuncts and finally to the special codes for place and personal names.

Directions to the program instruct the user to use the self-checking features in the program; and by the time the program had been completed the user could not only use the system with some facility, but had learned to use the program itself as a reference device in the later stages of learning, and when they began work on text samples. This program proved to be immensely successful, and simplified all later activity.

At an appropriate time in the development of this program for teaching edited Pin-yin romanization, we inserted an advertisement in three major papers to run for three days. The advertisement was for English typists with experience in romanization who wished to work part-time, and instructed them to send a vita sheet to a Post Office box. It further requested that anyone interested send in a romanization of the advertisement itself. This last feature was helpful and revealing in that it pre-selected those who had both typing and romanization skills, and were willing to put the effort into the transcription. We had more than two hundred replies, more than a hundred of which indicated that they had not understood the specifications. The rest were divided into three groups: a. A group which had submitted typed responses which minimally indicated access to a typewriter. All had good control of typing skills, and showed that they had mastered at least one romanization for Chinese. There were about thirty-five names in this group. b. A second lot which had poorer typing and romanization, and c. All others who had tried to fulfill the requirements, but simply could not perform effectively.

Those in the first group were rank-ordered on a scale which put primary emphasis on accuracy of typing and

correctness of romanizations. Neatness was judged only insofar as legibility was at issue. Clean erasures were accepted, but if numerous, served to downrank the individual.

The first twenty of these on the list were approached with a copy of the program developed to teach the Pin-yin romanization, and a sample to be coded. A few of these, on seeing the complexity of the task, indicated that they had no interest in such work, but almost all attempted to do one sample. They were told that the first samples would be paid at a premium rate, the second at a slightly lower rate, and all others at a flat rate, and that after the second sample was completed the rate would be reduced for evidence of error, so that a premium was put on accuracy. At the end of the first sample, it was clear that some were admirably suited to this work, and they were continued, and no contact was made with the others on the list. We ended up with about fifteen able transcribers whose accuracy rate was surprisingly high. It was these on whom we relied for all later transcriptions.

Coding Conventions

The texts are romanized using the Pin-yin system with a number representing the tone of the syllable incorporated at the end of each syllable, and 'v' representing ü. In addition to the digits one through four to indicate the four tones of Mandarin, and the use of five to indicate the tone of characters that are normally neutral, (qing1-yin1), four more 'tones' are also distinguished. Characters which are normally first tone, but in context are read as neutral are coded with '6'; neutral tones normally having the second tone are coded with '7', and similarly with the characters normally having the third and fourth tone. For purposes of retrieval and disambiguation, each tone from 1-4 is represented either by this number, or by this number plus five. In each case, the tone follows the romanization immediately with no intervening space.

Some grammatical and formatting information was also coded, using special identification marks. In considering frequency and word distribution, it is assumed that the graphic word rather than the single character in isolation is the unit to be considered. Therefore, words of more than one character are coded by connecting syllables with

BEST COPY AVAILABLE

a '-', with no spaces between the syllables. Such coding will permit the collection of data on homophones, but for recovery purposes will also serve to limit the number of homophomous 'words'. The standard usually applied is that of 'native-speaker intuition', and may result in some discrepancies in coding, although an effort has been made in the editing process to reduce these discrepencies.

Other connecting symbols are also used. When ordinarily 'free' syllables are formed into elements of words, such as the bu4 and the de7 in resultative compounds, the syllables in such a compound are connected by the plus sign. When the syllable is bound to the sentence as a whole rather than the word, the connecting symbol is the asterisk.

Personal and place names, as well as most of the format codes, have a single identifying symbol preceding the item, and a double symbol following the item. In each case the single or repeated symbol is both preceded and followed by a blank space to facilitate retrieval.

The following special codes are used for these purposes: Personal and place names: @...@@, Chapter headings: E...EE, Interlinear annotations: $...$$, Footnotes and other annotations: !...!!.

Since modern Chinese prose often cites foreign words or phrases which cannot be coded according to the conventions for Chinese material, material in this form is coded within parentheses as follows: The per cent symbol '%' preceded and followed by a blank, followed by a two-letter code for the language, followed by a period, which is then followed by a number indicating the number of words in the citation. The number is followed by a blank. An example: (% IT.6) represents the occurence of six words in Italian.

In addition, certain non-language data often occurs in Chinese. For this corpus, coding for figures, graphs, and charts and for equations and formulae is adequate. Each of these is coded by prefixing a double asterisk within parentheses both before and after the equation or figure, with a space before and after the double asterisk. The code for figures, graphs, or charts uses the capital letters 'FG', and that for equations or formulae is the capital letters 'FM'. Thus the presence of a figure or chart in the text is coded as (** FG **) and that for a formula or equation is (** FM **).

In order to save time in coding the original samples, and provide a psychological 'out' for a coder who might

BEST COPY AVAILABLE

not be willing to admit to inability to read some item, they were instructed to type out a sequence of seven x's when they were unsure of a romanization or reading for a character (XXXXXXX). Although attempts have been made to locate and give proper readings for such characters, this will remain the symbol for a character whose reading has not yet been determined for the specific context.

The first 70 columns of each card contain the text, column 71 is left blank, and columns 72 through 80 contain a location marker. The coding of the text in columns 1 through 70 is that described in the previous pages of this chapter. The location marker contains a line number in columns 72 through 75, the characters 'C1' in columns 76 and 77 (which identify the corpus and the first Chinese corpus, in analogy with the English 'E1' and the Russian 'R1', and a sample marker (78-80) consists of a letter corresponding to the classification of the sample as to genre, and a two-digit number (between 01 and 99) identifying the sample uniquely within that genre.

For purposes of the word distribution studies the graphic word is considered the basic unit and its integrity is never jeopardized. Therefore, if a word ends in column 70, column 1 of the next card is left blank and the next word is punched starting in column 2 without an intervening hyphen. If a word ends in column 69, column 70 is left blank and punching continues, starting in column 1 of the next card.

To facilitate the correction of proofread cards two conventions were adopted which are of particular interest to a programmer using the corpus: 1) a string of blanks greater than one is equivalent to a single blank and 2) the character is to be interpreted as "delete me and all the following blanks, if any occur". (The English uses * for this).

Records

A number of record-keeping errors have resulted in the loss of certain data for particular samples. In the manual, each of these is identified at the place at which the sample number occurs. Since these notifications are inserted in the text of the manual using the Imbed instruction available with NSCRIPT, alterations and updating will be possible as soon as the information becomes available. Users should request the latest

edition of the manual.

Manual

The manual was prepared by Misses Emily Calkins and Vicky Williams under the direction of Mr. Gerald Rubin of Brown University, using the text-editing program NSCRIPT. A text-editing program was chosen so that as corrections or additions to the manual are required they can be added to later editions with relative ease.

Since the manual is not intended to be part of the corpus itself, some of the conventions used in the manual differ from those in the corpus. Place names are romanized in Pin-yin, but without tones, except where the place is known by a more familiar romanization, or is a representation of a foreign word.

Thus, Xiang1-gang3 is 'Hongkong', and Luo2-si1-fu2 is 'Roosevelt'. Designations for streets and roads and other locations are translated rather than romanized, and where this might be ambiguous, as with the terms for 'village', the designation is followed by the romanization of the Chinese character in parentheses, as 'Panchiao Village' (zhen4), and 'Kuanghua New Village (tsun1).

Basic Technical Information

The Standard Chinese Corpus is available to interested scholars. It comprises one Magnetic Tape of about 1200 feet in length. The Corpus is available in either 7-track or 9-track mode.

The organization of the data on tape corresponds to the punched card format described above. The data is recorded on tape in card-image form, i.e., all cards (including correction cards and other incompletely filled cards) are reproduced in their entirety. However, for reasons of operating efficiency in processing the Corpus data, the card-image records are grouped on tape by a blocking factor of 40, so that the tape is composed of a series of fixed-length tape records each containing 3,200 characters (i.e., 40 cards). Since sample size and tape record length are independent of each other, the end of a sample does not necessarily coincide with the end of a tape record.

Availability

Both the manual and the Corpus are available from the Department of Linguistics, Brown University, Providence, Rhode Island 02912. The Corpus can be ordered in either 7-track or 9-track format and at several recording densities. Purchasers may send their own tapes to be copied onto for a service and handling charge of $50.00, or they may request that copied tapes be furnished at a charge of $75.00. Both these prices include one copy of the manual. Additional copies of the manual can be purchased for $10.00.

Acknowledgements

The success of this project is due in large measure to the efforts of two people, Mr. Gerald Rubin, my colleague in the Department of Linguistics at Brown, who wrote the programs necessary for efficient processing of the data, and Miss Ch'iu Pin, who selected many of the samples, interviewed the typist-transcribers, and kept the records after my departure from Taiwan. The typists and keypuchers are responsible for the accuracy of the product, although the errors are my responsibility. Mr. George L. Shelley, who acted as paymaster and who effected the liaison with all of the others involved in the project has given generously of his time and concern for more than two years in helping to bring the corpus to completion.

To all of these, all users should offer grateful thanks.